Fake News Spreader Identification in Twitter using Ensemble Modeling

Notebook for PAN at CLEF 2020

Ahmad Hashemi, Mohammad Reza Zarei, Mohammad Reza Moosavi, and Mohammad Taheri

Department of Computer Science and Engineering, Shiraz University, Shiraz, Iran
s.ahmad.hmi@gmail.com, mr.zarei@cse.shirazu.ac.ir,
smmosavi@shirazu.ac.ir, motaheri@shirazu.ac.ir

Abstract In this paper, we describe our participation in the author profiling task at PAN 2020. The task consists of detecting fake news spreaders on Twitter, based on a hundred selected tweets from their profile. In our approach, we utilized TF-IDF and word embeddings as text representation as well as taking advantage of statistical and implicit features. A combinational classification model is proposed to fuse the impact of all groups of features. The approach obtained highly competitive classification accuracies on both English (0.695) and Spanish (0.785) subsets of the task.

1 Introduction

With the rise of social media in the past decade, people have increasingly tended to seek out news from social media services rather than traditional news organizations due to the nature of social media platforms. For instance, receiving news from social media is often faster and cheaper than traditional news media, such as newspapers or television. Additionally, social media has the ability of sharing, commenting on, and discussing the news with friends or other users. However, despite these advantages, social media is exposed to fake news spread.

Fake news might be expected for any major event that captures people's imagination, but in some cases such as the COVID-19 pandemic, the spread of fake news offers unique challenges and dangers to the public. Due to the potential of spreading fake news to have serious negative impacts on individuals and society, investigating fake news is critical to help mitigate these negative effects.

Since users play an important role in the creation and propagation of fake news, it's valuable to identify users who spread fake news on social media. In the author profiling shared task at PAN 2020 [10], the focus is to identify fake news spreaders on social media using their provided textual contents as a first step towards preventing fake news from being propagated among online users. This year's task includes English and

Spanish languages. The provided dataset contains 300 users for each language with the textual content of 100 tweets per user. Also, an unseen test data containing 200 users for each language is available on the TIRA platform [9] for evaluating the final model. To increase the complexity, no metadata or other social network information is available.

The rest of this paper is structured as follows. In section 2, we provide an overview of the related work in author profiling and fake news detection on social media. In section 3, we discuss the overall design of the proposed model for the classification problem. Section 4 describes different groups of utilized features and explains their extraction procedure. Employed classifiers and the method of merging their decisions are illustrated in section 5. The experimental results are reported in section 6 and section 7 contains a brief conclusion.

2 Related Work

In this section, we discuss the related work from two aspects: fake news detection on social media and author profiling.

Fake News Detection on Social Media. Fake news on social media can be studied generally with respect to two perspectives: news content based and social context based. News content based approaches incorporate features from the textual or visual content of the news. For instance, in [4], the authors compare the language of false news with real news from an emotional perspective. Social context based approaches extract features from user profiles, post contents and social networks. In [16], the problem of understanding and exploiting user profiles on social media for fake news detection is studied. First, two groups of users are considered, the users who are more likely to trust and share fake news and the users who are more likely to share real news. Then, explicit and implicit profile features of each group are extracted and used for the fake news classification task.

Author Profiling. Author profiling has been undertaken as a shared task at PAN annually since 2013. Several aspects of author profiling in social media have been covered from 2013 to 2019 such as bot, age and gender detection. In various editions of the author profiling task at PAN, participants used a high variety of different features. Text representations such as character and word n-grams, bag-of-words (word unigrams), TF-IDF and word embeddings have been widely used as features. For instance, the best-performing team in textual classification at pan 2018 [2] proposed a linear Support Vector Machine (SVM) classifier, with different types of word and character n-grams as features [12]. Some teams also used stylistic features as in [6] that the authors aggregated statistics in addition to using term occurrences. Although the impact of emotions on author profiling has been investigated before [11], the use of other implicit features such as personality dimensions has been neglected in author profiling tasks.

3 Overview of The Proposed Model

The proposed model consists of three different components, each designed to utilize a group of particular features. For each component, we manipulate a group of features: (i) features extracted from word embeddings; (ii) extracted features using TF-IDF; (iii)

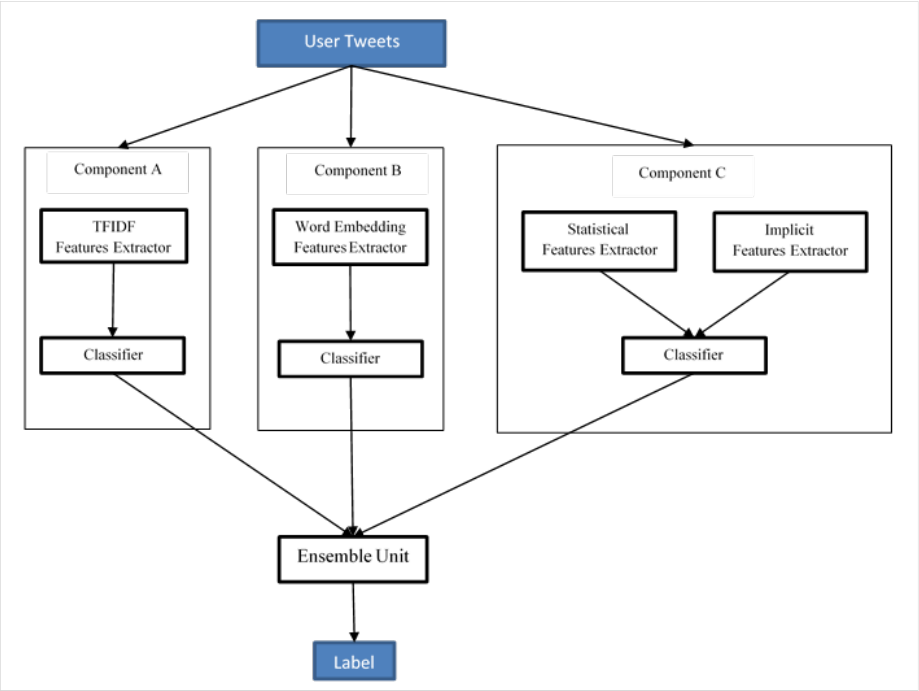

Figure 1. Overview of the proposed model

the combination of statistical and implicit features. Each component of the model processes a set of features with a distinct classifier. Then, the results of the classifiers are aggregated with an ensemble method. The proposed model is illustrated in Figure 1.

As different groups of features have distinguished aspects, presenting a model to utilize various types of features could enhance the final performance. The details of the proposed model are described in the following sections.

4 Feature Extraction

4.1 Preprocessing

For all groups of features in both languages, we utilized the TweetTokenizer module from the Natural Language Toolkit (NLTK) package [1] of Python as well as omitting retweet tags, hashtags, URLs and user tags. For TF-IDF features, in addition to eliminating punctuations, numbers and stop words, we generated the root of words for English and Spanish with Porter and Snowball stemmers, respectively.

4.2 Feature Groups

Word Embeddings To extract word embedding vectors, we utilized medium-size pretrained models of English and Spanish languages available in the Spacy package [5].

The sources used in the English model for training data are OntoNotes 5[1] and GloVe Common Crawl[2] and the Spanish model utilizes UD Spanish AnCora v2.5[3], WikiNER[4], OSCAR (Common Crawl)[5] and Wikipedia (20200301)[6]. Then, each user is represented as the weighted average of word embeddings for all the words available in the user's tweets. The frequency of the words in the user's tweets is used as the weight factor.

Term Frequency - Inverse Document Frequency We used TF-IDF which is a common technique in Information Retrieval and Text Mining, to create another set of features for each user. The tweets of each user are concatenated to construct an individual document and then, its TF-IDF feature vector is used as the features of the user.

Statistical Features Since no metadata is available in the dataset, we attempted to extract such features from the textual content by calculating the following statistics for each user:

- Fraction of retweets (tweets starting with "RT")
- Average number of mentions per tweet
- Average number of URLs per tweet
- Average number of hashtags per tweet
- Average tweet length

These features are extracted for both English and Spanish languages.

Implicit Features Implicit features are a group of user profile features that are not directly available and here are inferred from the textual content of users' historical tweets. We focused on those implicit features that are commonly utilized to better describe user profiles. Our explored implicit profile features for the English language are: age, gender, personality dimensions and emotional signals. We also extract emotional signals for the Spanish language.

All implicit features are extracted using lexicon-based approaches. For each group of features we used a state-of-the-art weighted lexicon. Despite each lexicon has its own approach [7], a weighted lexicon can be generally applied as follows:

$$score(profile, lex) = \sum_{word \in lex} w_{lex}(word) \ * \ \frac{freq(word, profile)}{freq(*, profile)} \qquad (1)$$

where $profile$ is the concatenation of all tweets belonging to a user, $w_{lex}(word)$ is the lexicon weight for the $word$, $freq(word, profile)$ is frequency of the word in the profile, and $freq(*, profile)$ is the total number of words used in the user's tweets.

[1] https://catalog.ldc.upenn.edu/LDC2013T19
[2] https://nlp.stanford.edu/projects/glove/
[3] https://github.com/UniversalDependencies/UD_Spanish-AnCora
[4] https://figshare.com/articles/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500
[5] https://oscar-corpus.com/
[6] https://dumps.wikimedia.org/
[7] The exact usage of each lexicon can be found in its related reference.

Gender and Age. Age and gender are widely used features to better describe users [15]. We extract these features by applying a state-of-the-art approach [14] on the user's Tweets.

Personality dimensions. Personality can be defined as the characteristic sets of behaviors and cognitions that distinguishes an individual from others. We attempted to extract the popular Big Five personality dimensions, also known as the five-factor model(FFM), which is a suggested taxonomy to classify the human personality into five dimensions: **Openness** (inventive/curious vs. consistent/cautious), **conscientiousness** (efficient/organized vs. extravagant/careless) **Extraversion** (outgoing/energetic vs. solitary/reserved), **Agreeableness** (friendly/compassionate vs. challenging/callous), **Neuroticism** (sensitive/nervous vs. resilient/confident). To predict users' personalities, we applied an open-vocabulary approach [15].

Emotional signals. Since the impact of emotions on previous author profiling tasks has been revealed [11], we decided to utilize emotional signals as another group of implicit features. The emotional signals that appeared in each user's tweets were extracted as the user's features. To extract the emotional signals, we used two lexicon-based approaches, one for English [13], and another for Spanish [3]. Finally, we extracted 8 emotions for the English subset of the task (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and 6 emotions for the Spanish subset (anger, fear, joy, repulsion, sadness, surprise) based on the available lexicons.

5 Classifiers

For each group of features, we selected the classifier with the best performance among Logistic Regression, Random Forest and support Vector Machine (SVM) with linear kernel. The best results were obtained with Random Forest on all features of both languages with an exception of SVM with linear kernel on TF-IDF features of the Spanish language. To merge the final decisions of distinct classifiers more effectively, we considered soft classifiers to obtain the confidence of each model for their predictions, instead of using just the predicted labels. We used scikit-learn's implementation [7] for our models.

The overall confidence of the method for an arbitrary user u is calculated as:

$$c_{out}(u) = \alpha c_1(u) + \ \beta c_2(u) + \ \gamma c_3(u) \tag{2}$$

where $\alpha+\beta+\ \gamma = 1$. $c_1(u)$, $c_2(u)$ and $c_3(u)$ are the confidence of the classifiers for TF-IDF, Word Embeddings and implicit+statistical features, respectively. The contribution of the classifiers in final confidence is controlled using the weight parameters α, β and γ, which their best values are determined using a greedy search on the train set.

Finally, the label of the user u is determined as:

$$y\,(u) = \begin{cases} 0 & if\ c_{out}\,(u) \leq 0.5 \\ 1 & if\ c_{out}\,(u) > 0.5 \end{cases} \tag{3}$$

6 Experimental Result

To obtain the best performance of our model, we evaluated the accuracy with various settings by applying stratified 10-fold cross-validation on the task's train set.

As mentioned before, for each component, the classifier with the best performance is selected among Logistic Regression, Random Forest and SVM with linear kernel. Table 1 shows the cross-validation mean accuracy scores for each group of features with these classifiers. All numbers are expressed as percentages. For TF-IDF features on Spanish Language, SVM obtained the best result while for other features on both Spanish and English languages, the best performance belonged to Random Forest.

Table 1. Accuracy scores of 10-fold cross-validation to select the best performing classifier for each group of features

Feature group	Language	SVM	Random Forest	Logistic Regression
Statistical+Implicit	English	57.6	**69**	49.6
TF-IDF	English	68.3	**70.3**	68.3
Embeddings	English	67.6	**71.3**	67.6
Statistical+Implicit	Spanish	72.6	**73**	56
TF-IDF	Spanish	**82**	80	81.6
Embeddings	Spanish	74	**76.3**	76

As described in the previous section, we merged the classifiers' decisions using equation 2. The weight parameters α, β and γ are optimized with grid search according to cross-validation mean accuracy. The best weight parameters are illustrated in Table 2. In the English language subtask, the word embeddings component owned the highest contribution with a 0.45 weight while the second and third ranks belonged to the components of Statistical+Implicit and TF-IDF, respectively. In the Spanish language subtask, the TF-IDF component achieved the highest rank and the minimum weight belonged to the word embeddings component.

Table 2. Determined weight parameters for merging the individual classifiers

Language	TF-IDF (α)	Embeddings (β)	Statistical+Implicit (γ)
English	0.15	0.45	0.4
Spanish	0.65	0.1	0.25

Table 3 shows the cross-validation scores for the individual components and the final model. As it is clear, utilizing a combination of the components has improved the accuracy of both English and Spanish languages compared to using each component, individually.

After achieving the best model in our local evaluations, we trained the model on the whole train set and tested it on the official unseen test set provided for the task, on the TIRA platform [8]. The results are shown in Table 4.

Table 3. Cross-validation scores obtained on different components

Model	Mean Accuracy (EN)	Mean Accuracy (ES)
TF-IDF	70.3	82
Embeddings	71.3	76.3
Implicit+Explicit	69	73
Combinational model (final model)	**74.6**	**82.9**

Table 4. Accuracy scores obtained on the local evaluation and the official test set

Language	Cross-validation	Test set
English	74.6	69.5
Spanish	82.9	78.5

7 Conclusion

In this notebook, a combinational model is proposed for the author profiling task at PAN 2020. The purpose of this model is to utilize an ensemble of the features: TF-IDF, word embeddings, statistical and implicit features to benefit from advantages of the features altogether and enhance the overall performance. An individual component with its specific classifier is considered for each group of features and in the test phase, the decision of all components are merged using a fusion formula. For all groups of the features except TF-IDF features of the Spanish language, Random Forest is used due to better performance compared to SVM and Logistic Regression. For TF-IDF features of the Spanish language, SVM with linear kernel is utilized. Regarding the results on the train set of the task, our combinational model outperforms each individual component which is equal to using just a group of features with its own best performing classifier.

For future work, utilizing more features especially the implicit ones could be considered. Furthermore, proposing a learning scheme for the fusion component could lead to more efficient and improved optimization of weight parameters of the utilized components.

References

1. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc (2009)
2. Daneshvar, S., Inkpen, D.: Gender Identification in Twitter using N-grams and LSA Notebook for PAN at CLEF 2018. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J.Y., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018) (2018)
3. Díaz Rangel, C., Guerra, S.: Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. Onomazein **29**, 31–46 (2014). https://doi.org/10.7764/onomazein.29.5
4. Ghanem, B., Rosso, P., Rangel, F.: An Emotional Analysis of False Information in Social Media and News Articles. ACM Transactions on Internet Technology **20**(2), 17 (aug 2019)

5. Honnibal, M., Montani, I.: spaCy2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. (2017). https://doi.org/10.5281/zenodo.1212304
6. Johansson, F.: Supervised Classification of Twitter Accounts Based on Textual Content of Tweets Notebook for PAN at CLEF 2019. In: Cappellato, L., Ferro, N., Lasada, D.E., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers (2019), CEUR-WS.org
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**(85), 2825–2830 (2011)
8. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF. Springer (2019)
9. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World. Springer (Sep 2019)
10. Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2020)
11. Rangel, F., Rosso, P.: On the impact of emotions on author profiling. Information Processing and Management **52**(1), 73–92 (jan 2016). https://doi.org/10.1016/j.ipm.2015.06.003
12. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. Working Notes Papers of the CLEF (2018)
13. Saif, M.: Word Affect Intensities. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018)
14. Sap, M., Park, G., Eichstaedt, J.C., Kern, M.L., Stillwell, D., Kosinski, M., Ungar, L.H., Schwartz, H.A.: Developing Age and Gender Predictive Lexica over Social Media. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1146–1151. Association for Computational Linguistics, Doha, Qatar (2014). https://doi.org/10.3115/V1/D14-1121
15. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H.: Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE **8**(9), e73791 (sep 2013). https://doi.org/10.1371/journal.pone.0073791
16. Shu, K., Zhou, X., Wang, S., Zafarani, R., Liu, H.: The role of user profiles for fake news detection. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019. pp. 436–439. Association for Computing Machinery, Inc (aug 2019). https://doi.org/10.1145/3341161.3342927